ASSOCIATION BETWEEN MODULATION SPECTRUM AND SPEECH INTELLIGIBILITY OF SYLLABLE-TIMED LANGUAGES

Guangting Mai, Gang Peng & William S-Y Wang

Language Engineering Laboratory, The Chinese University of Hong Kong, Hong Kong
kongtingmak@gmail.com; gpeng@ee.cuhk.edu.hk; wsywang@ee.cuhk.edu.hk

ABSTRACT

Previous studies showed that both amplitude [1, 6] and phase [4] of the Modulation Spectrum (MS) of speech waveforms play an important role in preserving intelligibility in stress-timed languages like English. In the current study, association between MS and speech intelligibility of spoken sentences in Mandarin and Cantonese which are typical syllable-timed languages [7, 8], is investigated. The manipulation of “local time reversal” on speech waveforms was employed for each sentence. Speech identification accuracies were calculated and MS analysis was implemented. It is found that both amplitude and phase of the MS components at the corresponding syllabic rates of the spoken sentences are contributing to speech identification. We suggest that this work will help us understand more about the relation between speech intelligibility and speech acoustics, especially for syllable-timed languages.

Keywords: modulation spectrum, modulation index, phase shift, syllable-timing, intelligibility

1. INTRODUCTION

Speech signals are highly tolerant with distortions. For instance, speech can be highly intelligible even when the temporal or spectral information are greatly reduced [1, 9, 10]. However, the question of what are the essential cues for speech intelligibility is still unresolved. In 1980s, Houtgast and Steeneken [6] showed that the low-frequency temporal envelopes at 3 to 8 Hz are important for intelligibility of English spoken sentences. More recently, the importance of slowly fluctuated temporal envelopes continued to be highlighted: Shannon, *et al.* [10] and Dorman, *et al.* [3] applied “noise-vocoder” and found that when preserving the speech envelopes in only a few frequency channels, participants could still have very good or even perfect understanding of spoken sentences.

The findings that temporal envelopes are essential for understanding speech signals gradually led to investigations on the relationship between Modulation Spectrum (MS) and speech intelligibility. MS was first demonstrated in [6], which refers to the power spectrum of signal’s temporal envelopes and depicts the distribution of energy in the amplitude fluctuations across frequencies. A recent study by Arai and Greenberg [1] applied manipulations of “cross-channel spectral asynchrony” (desynchronizes spectral information across different frequency channels) and found that attenuation in the amplitude of the MS at 3 to 6 Hz is highly associated with intelligibility reduction. Another acoustic analysis on English speech by the same authors showed that not only the amplitude but also the phase of the MS are important for speech intelligibility [4].

In our present study, we further investigated the association between the MS (including both its amplitude and phase) and speech intelligibility of spoken sentences in Mandarin and Cantonese, which are considered to be typical syllable-timed languages [7, 8]. Previous [5] and our present studies reveal the different MS energy distributions across modulation frequencies [1] for syllable-timed (such as Mandarin and Cantonese) and stressed-timed languages (such as English). We here followed the study by Greenberg and Arai [4] employing the “locally time-reversed speech” [9] (slice speech signals into segments of *equal* duration and then time-reverse each segment), by which syllabic rates of the original stimuli were carefully controlled.

2. BEHAVIORAL EXPERIMENTS

2.1. Stimuli and tasks

300 different Mandarin and Cantonese “Semantically Unpredictable Sentences” (SUSs, 150 SUSs for each of the two languages) were created and naturally pronounced at 4, 6 and 8 Hz syllabic rates by a single male Mandarin-Cantonese bilingual speaker sampled at 44.1 kHz. The SUSs are sentences that are syntactically acceptable but semantically anomalous [2], and

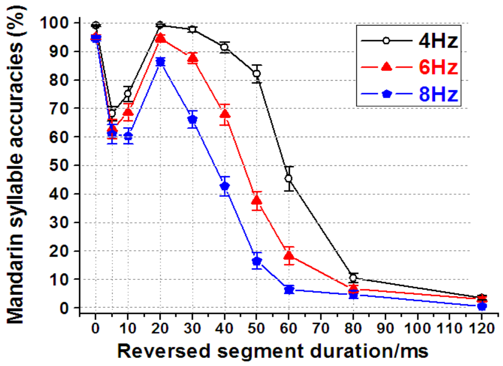
each SUS is constituted of several disyllabic words like “工作庆祝广泛的特征” (“Jobs celebrate the vast traits”). The SUSs were then locally time-reversed with segments of various durations ranging from 5 to 120 ms (specifically, 5, 10, 20, 30, 40, 50, 60, 80 and 120 ms). In addition, original unreversed SUSs were also included in the experiment (designated as “0 ms reversal”).

18 naive Mandarin-Cantonese bilingual subjects recruited from Guangdong Province in Mainland China (7 male, 11 female, aging from 21 to 26 years old) were instructed to listen to both Mandarin and Cantonese SUSs (each sentence played once only) and write down the words or syllables they heard.

2.2. Behavioral results

Syllable identification accuracies were calculated for different reversed segment duration conditions (Fig. 1). Fig. 1 shows the result for Mandarin SUSs and the result for Cantonese shares a similar pattern (shown in supplementary file *image file 1*): (1) modest attenuation in accuracies relative to the original unreversed conditions (0 ms reversal) happens at 20 to 50 ms reversed segment durations for 4-Hz, 20 to 30 ms for 6-Hz and 20 ms for 8-Hz syllabic rate SUSs, respectively, keeping the syllable accuracies above 80%; (2) more acute decline in accuracies is found for longer reversed segment durations (e.g., at 120 ms reversal for 4-Hz syllabic rate SUSs, the accuracy is below 10%); (3) more interestingly, at reversed durations shorter than 20 ms (namely 5 and 10 ms), there exists significantly abrupt decrease in accuracies for all the three syllabic rate SUSs (lower than 80%), which has never been found in any other studies.

Figure 1: Mandarin syllable accuracies as a function of reversed segment duration, where the black, red and blue curves stand for 4-, 6-Hz and 8-Hz syllabic rate SUSs, respectively. Error bars represent Standard Errors across all the 18 subjects. The result for Cantonese is shown in *image file 1*.



3. ACOUSTIC ANALYSIS ON THE MODULATION SPECTRUM

3.1. Acoustic analysis process

3.1.1. Frequency channels for analysis

60 SUSs in the experiment were randomly chosen for MS analysis (30 for each of the languages). Each sentence was band-pass filtered into 12 frequency bands in logarithmic scale referenced to the study by Dorman, *et al.* [3] which revealed that preserving the temporal envelopes in 8 bands from 300 to 5500 Hz is sufficient enough for English spoken sentence identification. Here we applied the same 8 bands and furthermore added another 4 bands from 71 to 300 Hz in accordance with the same logarithmic scale and conducted the analysis in each band. The intention of adding these 4 bands is to cover the range of the first and second harmonics in the speech stimuli (from around 90 to 300 Hz), which are important for lexical tone recognition. The frequency ranges of each band are shown in Table 1.

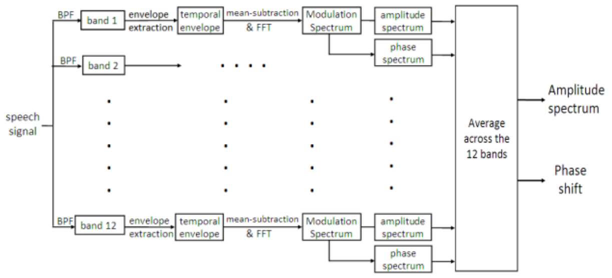
Table 1: Frequency ranges and bandwidths of the 12 frequency channels for MS analysis.

Channel No.	Frequency range (Hz)	Bandwidth (Hz)
1	71 ~ 103	32
2	103 ~ 146	43
3	146 ~ 210	64
4	210 ~ 300	90
5	300 ~ 432	132
6	432 ~ 621	189
7	621 ~ 893	272
8	893 ~ 1284	391
9	1284 ~ 1847	563
10	1847 ~ 2657	810
11	2657 ~ 3823	1166
12	3823 ~ 5500	1677

3.1.2. Obtaining the Modulation Spectrums

To obtain the MS (both its amplitude and phase components) in each frequency band, the following steps were manipulated for each sentence (Fig. 2).

Figure 2: Procedures of MS analysis in each SUS. Abbreviation: BPF for Band-Pass Filtering, FFT for Fast Fourier Transform.



Step (1): envelope extraction. Signals in each band were half-wave rectified and low-pass filtered at 30 Hz to obtain the temporal envelope. *Step (2)*: the temporal envelope was then subtracted from its own average value and subsequently Fast Fourier Transformed (FFT) to obtain the MS. *Step (3)*: the obtained MS in Step (2) was then decomposed into amplitude and phase spectrum in each frequency band. The MS amplitude is represented as Modulation Index (MI) [6] which is equivalent to the *normalized* amplitude value.

$$MI = \frac{A}{A_0} \qquad (1)$$

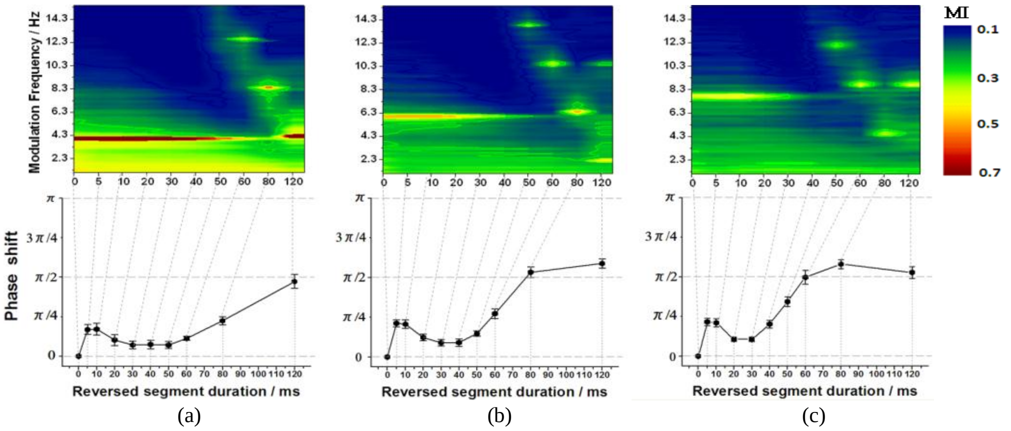
where A refers to the MS amplitude value and A_0 denotes the average value of the temporal envelope extracted in Step (1) (both in the unit of sound pressure *Pascal*). The phase of MS is represented as phase shifts relative to the original SUSs at the modulation frequencies in the neighborhood of the corresponding syllabic rates. The reason for this manipulation is that the energies of the MS are centered at the frequency components same as the syllabic rates (see details in Section 3.2).

Same manipulations were conducted in each of the SUSs and the results were finally grandly averaged.

3.2. Results

One of the results of the analysis is shown in Fig. 3, illustrating the amplitude spectrums and phase shifts of the MS of Mandarin SUSs and how they associate with the intelligibility data shown in Fig. 1 (detailed discussions as below). Similar results were obtained for Cantonese SUSs shown in the supplementary file *image file 2*.

Figure 3: MS analysis results for *Mandarin* SUSs averaged across the 12 frequency channels. Upper panels: amplitude spectrums. The color bar represents the magnitude of the Modulation Index (MI). Lower panels: phase shifts with error bars of Standard Errors across all the 12 frequency channels. (a) 4-Hz syllabic rate SUSs. (b) 6-Hz syllabic rate SUSs. (c) 8-Hz syllabic rate SUSs.



Upper panels in Fig. 3 show the amplitude spectrum as a function of modulation frequency (vertical axis) and reversed sement duration (horizontal axis) in all the three syllabic rate SUSs ((a), (b) and (c) stand for 4-, 6- and 8-Hz, respectively). Lower panels display the corresponding phase shifts at the frequencies in the neighborhood of the syllabic rates (3.5 to 4.5 Hz for 4-Hz, 5.5 to 6.5 Hz for 6-Hz and 7.5 to 8.5 for 8-Hz syllabic rate SUSs).

Take 4-Hz syllabic rate SUSs for instance (Fig. 3(a)), the MS energies are superiorly congregated at around 4-Hz in the original signals (at 0 ms reversal, as shown in the upper panel), reflecting the syllable-timed characteristics of Mandarin continuous speech. There are little changes in the MI value at 4 Hz frequency components at 20 to 40 ms reversals relative to the original signals, which is aligned with the behavioral performances that the syllable accuracies are still kept above 90% at these reversals. The gradient of the MI attenuation at 4 Hz from 50 to 80 ms reversals also matches well with the intelligibility reduction at these reversals. On the other hand, however, the MI values do not correlate with the accuracy descents at 5, 10 and 120 ms reversed durations.

Phase shifts relative to the original signals at the frequencies within 3.5 to 4.5 Hz (the lower panel of Fig. 3(a)) seem to additionally contribute to intelligibility. At 5 and 10 ms reversals, the phase shifts are significantly larger than the ones at 20 to 40 ms reversals ($p < 0.05$, according to One-way ANOVA comparing 5 and 10 ms with 20 to 40 ms reversals across the 12 frequency bands).

This result is associated with the behavioral performance that syllable accuracies are apparently lower at 5 and 10 ms reversals than at 20 to 40 ms reversals (Fig. 1(a)). At 120 ms reversal, the phase shift is up to π/2, revealing the orthogonal relation between the temporal envelopes with 120 ms segment-reversal and envelopes extracted from the original signals, which means little correlations exist between the two. We suggest this is probably the reason that almost zero intelligibility is found at this reversal condition.

Similar associations between the MS (MI values and phase shifts) and syllable accuracies are also observed in the results of 6-Hz (Fig. 3(b)) and 8-Hz (Fig 3(c)) syllabic rate SUSs.

4. GENERAL DISCUSSION

In the current study, two parameters were carefully set. Firstly, the syllabic rates of the Mandarin and Cantonese speech stimuli were well controlled at 4, 6 and 8 Hz, respectively. Secondly, frequency channels selected for the MS analysis were referenced to the previous study on English spoken sentences [3] and further took account of the tonal characteristics of Mandarin and Cantonese (detailed descriptions in Section 3.1.1).

Dramatic attenuation in speech identification accuracies in both Mandarin and Cantonese spoken sentences are accompanied by the distortions of the Modulation Spectrum, either on its amplitude or phase of the frequency components at the corresponding syllabic rates of the sentences. This happens at both long reversed segment durations (e.g. >50 ms for 4-Hz-syllabic-rate SUSs) and very short durations (5 and 10 ms) for all the three syllabic rate SUSs.

We suggest that such result is correlated with the syllable-timed features of Mandarin and Cantonese spoken sentences. It has been previously found that the more variations in the duration of syllables, the more dispersive the MS energy distribution is observed in a spoken language [5]. Syllable-timed languages tend to have relatively steadier syllable durations, while in contrast, durations between stressed and unstressed syllables are of great disparity in stress-timed languages like English [5]. This results in more concentrated MS energy distributions for syllable-timed than stress-timed languages. In the present study, we have found that due to the syllable timing, the MS energies of both Mandarin and Cantonese SUSs are superiorly concentrated at the modulation frequencies of the corresponding syllabic rates (Fig. 3). Thus, other than other MS frequency components, the component at the syllabic rate of the sentence is likely to play a major and more important role in speech intelligibility for syllable-timed than stress-timed languages, which is also the reason for our current focus on the MS components at the corresponding syllabic rates of the sentences. However, this is not to say there should be no contributions to Mandarin or Cantonese speech intelligibility made by other frequency components, the importance of which we should further study in our future work.

In summary, the current study investigated the association between Modulation Spectrum and speech intelligibility of Mandarin and Cantonese spoken sentences with local time-reversal. This study will help us understand more about the relationship between speech intelligibility and speech acoustics, especially for syllable-timed languages.

5. REFERENCES

[1] Arai, T., Greenberg, S. 1998. Speech intelligibility in the presence of cross-channel spectral asynchrony. *IEEE ICASSP* Seattle, 933-936.

[2] Beno î, C., Grice, M., Hazan, V. 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable Sentences. *Speech Communication* 18, 381-392.

[3] Dorman, M., Loizou, P., Rainey, D. 1997. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *J. Acoust. Soc. Am.* 24, 175-184.

[4] Greenberg, S., Arai, T. 2001. The relation between speech intelligibility and the complex modulation spectrum. *Proc. 7th EuroSpeech* Aalborg, 473-476.

[5] Greenberg, S., Arai, T., Grant, K.W. 2006. The role of temporal dynamics in understanding spoken language. In Divenyi, P., Vicsi, K., Meyer, G. (eds.), *Dynamics of Speech Production and Perception*. Amsterdam: IOS Press, 171-190.

[6] Houtgast, T., Steeneken, H. 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.* 77, 1069-1077.

[7] Lin, H., Wang, Q. 2007. Mandarin rhythm: An acoustic study. *Journal of Chinese Linguistics and Computing* 17, 127-140.

[8] Mok, P.K., Dellwo, V. 2008. Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. *Speech Prosody 2008* Campinas, Brazil, 423-426.

[9] Saberi, K., Perrott, D. 1999. Cognitive restoration of reversed speech. *Nature* 398, 760.

[10] Shannon, R., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M. 1995. Speech recognition with primarily temporal cues. *Science* 270(5234), 303-304.

[1] The term “modulation frequency” refers to the frequency (in Hz) of the slowly fluctuated temporal envelopes below 16 Hz [6], as shown in Fig. 3.